INF 558 – Spring 2019

Homework #5: RDF

Due: 3/27, Wednesday
100 points
(TA handling this homework: Binh Vu binhlvu@usc.edu)

Open graph protocol (namespace "og") is commonly used to describe a web page as a node in a graph (e.g., social network graph or graph of objects). It consists of a predefined list of properties for describing the properties of web pages. Example properties are: og:title, og:type, og:url, an og:image.

The properties of objects are represented in attributes of <meta> tags (like in RDFa). For example, `<meta property="og:type" content="article" />`.
This says that the object has a property called "og:type" and its value is "article".

See the web site: http://ogp.me/ for details on these properties and many others.

In this homework, you are provided a collection of pages from allrecipes.com ("webpages" folder), each containing the structured data in <meta> tags described using the open graph protocol. Your tasks are as follows:

1. [80 points] Implement a Python 3 program "extract.py" that extracts property names and values for each page in the collection of pages. It should store the extracted information in a text file in the JSON-LD format (<output_jsonld_file>).

   Execution format: `python extract.py <input_dir> <output_jsonld_file>`

   Where <input_dir> stores the collection of pages. You may use URL of page (e.g., https://www.allrecipes.com/recipe/147988/banana-bran-zucchini-bread/) as the value of its "@id" property. The URLs can be found in "index.json" file in the "webpages" folder

   Example: `python extract.py ./webpages ./output.jsonld`

2. [20 points] Implement a Java program that uses Apache Jena to convert the JSON-LD data created in the first task into Turtle format. It should also report, for each page, the number of properties in the page. You should use the provided skeleton in "jsonld2ttl" folder (maven project) for this task. You can import this project into NetBeans or any preferable IDE.

   The program will be compiled and run as follow:

   ```
   mvn clean package
   ```

```
java -cp target/jsonld2ttl-1.0-SNAPSHOT.jar edu.usc.inf558.App
<input_jsonld_file> <output_ttl_file>
```

Submission instructions:

You must submit the following files in a single .zip archive, which is named Firstname_Lastname_hw5.zip, to Blackboard. Your homework **will not be graded** if your submission is in the wrong format.

- extract.py: your extraction program
- requirements.txt: (optional) list of required python libraries for your **extract.py** program. This will be installed using the command: `pip install -r requirements.txt`. The pre-installed libraries are anaconda3 and beautifulsoup4.
- jsonld2ttl: directory that contains your maven project for task 2.
- allrecipes.jsonld: the jsonld file you generated in task 1.
- allrecipes.ttl: the turtle file you generated in task 2.

You can validate your submission using this docker command

```
docker run --rm -v <directory_that_contains_your_submission>:/submission
toan2/inf558-2019:hw5
```